# Exploratory Analysis For Online News Popularity - A deep insight analysis

In this project, the goal is to explore the dataset given and be able to find critical insights that can be used to influence potential article popularity. Also, machine learning models were built to be able to predict the popularity of a given article.

The process followed is highlighted below:

* Data Cleaning - Noise detection and removal
* Subjective analysis - Using our intuition to evaluate a data variable/feature and decide whether a variable influences the popularity of the article or not.
* Quantitative Analysis - How correct is our intuition? Here we carry our several analysis to accept or debunk our initial hypothesis
* Normal Distribution Observation on the dataset
* Feature Selection and Evaluation
* Machine Learning Classification
* Summary and Conclusion.

## Open Book - Ideas to consider

Some ideas to consider

* What is the effect of number of images/number of videos on the article's popularity?
* Is there a relationship between the number of words in the content and/or  
  number of words in the title in the article popularity
* Is their a concrete relationship between average length of words in the content to the popularity
* Create a grading/rank for the popularity: Excellent, Good, Okay, Poor, Very Poor
* How is the ranking of the channels in regards to the shares popularity
* What data channel has the most popularity and what feature in that particular data channel contributes towards that assertion. Is this also observed in the other data channels?
* What about the effects of the Worst, Best, and Avg keywords
* Is there an influence on the min, max or avg shares on each article referenced. Does the number of shares in those referenced articles also influence the number of shares in the main article?
* At what point in the weekend do people share articles the most? Can that mean people read those articles the most on those days?
* What is the effect of LDA analysis on the article popularity
* Does article with more text sentiment influence the popularity
* What is the relationship between the text sentiment and the article publish day? Are there more sentiments on a particular day?
* How about the influence of positive/negative words in the text sentiment and popularity. Do people favour positively worded articles? What is the ratio of positive to negative word articles in the dataset? Are the mutually balanced and can we make a judgment based on that alone?
* What about the effect of subjectivity on the title and test in the popularity

## Grading the Shares

* Exceptional = Top 95%
* Excellent = Top 90%
* Very Good = Top 80%
* Good = Top 60%
* Average = Top 50%
* Poor = Top 35%
* Very Poor = Rest

# Making Recommendations For Good Articles

* n\_tokens\_content should be less than 1500 words. The lesser the better.
* n\_tokens\_title should be between 6 - 17 words.
* n\_unique\_tokens should be between 0.3 - 0.8
* num\_hrefs is between 1 and 40 reference links
* num\_imgs should between 1 - 40 images
* num\_videos should be between 0 - 25 videos. The higher the lower the odds.
* average\_token\_length should be between 4 - 6
* The number of keywords in the metadata really influences the shares to a margin. The higher the value the better the shares chances. A value upward of 5 is recommended.
* Here, it can be seen that the best articles with highest share popularity belongs to the "Others" channel. For a more concrete channel, The "Business" and "Entertainment" channels are great for the best popularity. Coming in third position will be the "World" and "Tech".
* Best popular articles are usually posted on Mondays and Wednesday (and a bit of Tuesdays). Sundays and Saturdays (Weekends generally) are the worst days to publish an article.
* Articles that talks about current trending are better for shares
* From the scatterplot below, it can be seen than good articles will generally tend to have n\_tokens\_content less than 2000 and greater than 100 words

# 

# Feature Extraction & Selection

* Here we will be extracting some of the best features we observed out from the data. The below criteria will be considered:
  + Feature selection based on best hypothesis observed
  + Feature Selection on the whole dataset
  + Feature selection using fisher discriminal analysis
  + Feature selection based on the best hypothesis observed but with a normal distribution (log transformation)
  + Feature selection using fisher discriminate analysis on normal distribution dataset
  + Feature Selection on the whole dataset - Normal Distribution

### Variables of our features selection are listed below:

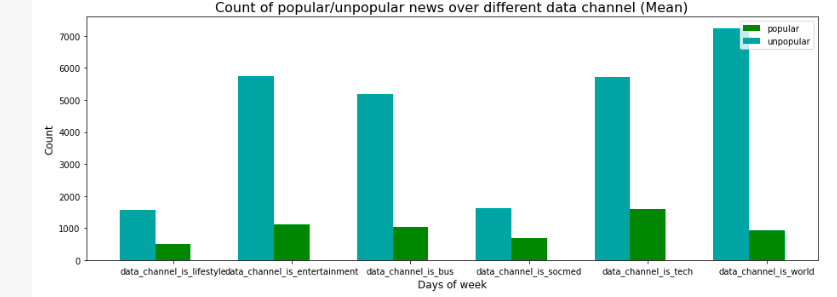
* Feature selection based on best hypothesis observed - **data\_feature1**
* Feature Selection on the whole dataset - **data\_feature2**
* Feature selection using fisher discriminal analysis - **data\_feature\_fisher**
* Feature selection based on the best hypothesis observed but with a normal distribution (log transformation) - **data\_feature1\_normal**
* Feature selection using fisher discriminate analysis on normal distribution dataset - **data\_feature\_fisher\_normal**
* Feature Selection on the whole dataset - **data\_feature2\_normal**

The KNN model which gave the best accuracy of 49.11% was based on using the all the data-set feature and number of neighbors of 71. Although this was the best accuracy discovered, there wasn’t much difference with the other models. For example, using all the features gave an accuracy less than 1 % of the highest accuracy observed.

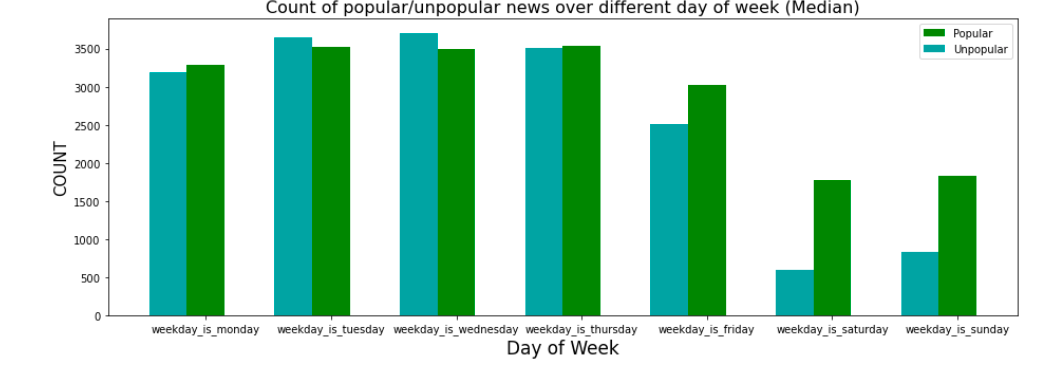
Random Forest has the best result for this classification task reaching an accuracy of 51.4%. Due to the nature of Random forest being able to set different numbers of decision trees, features, tree depth, splitting criteria, and others it tends to require a lot of parameter tuning.

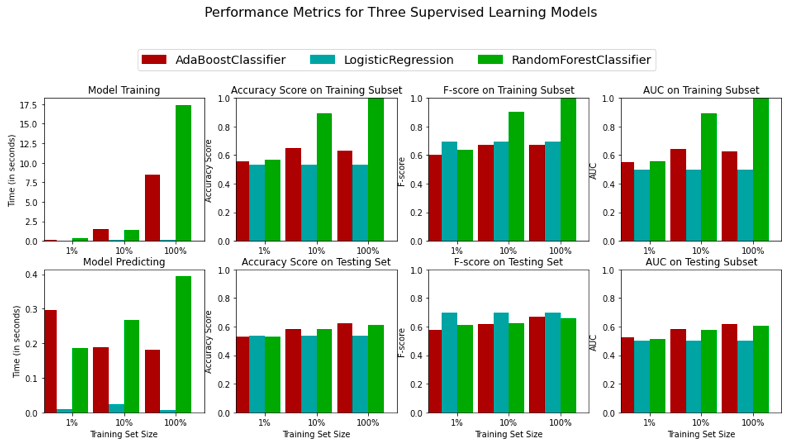
The maximum accuracy observed with SVM was 50,58% . An observation with SVM is that training starts becoming increasing as the number of polynomial degrees increases, training examples increases, C value rises and also the number of features increase which basically makes the model become more complex to draw an hyperplane for separating the classes.

# GRAPHS

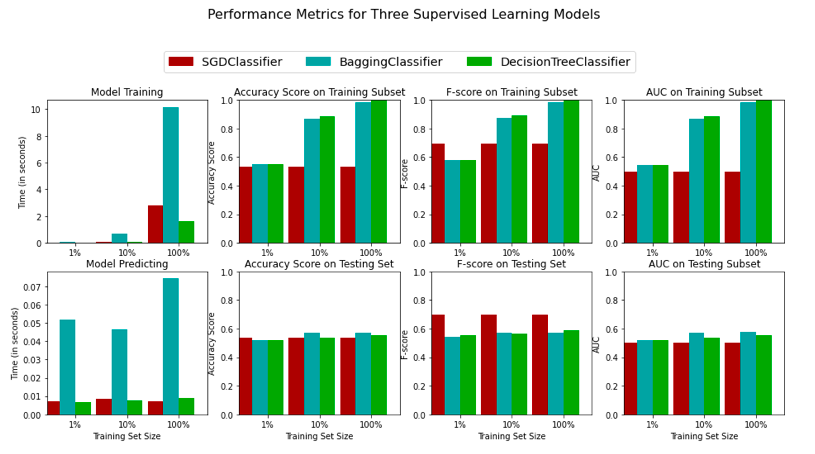


# 

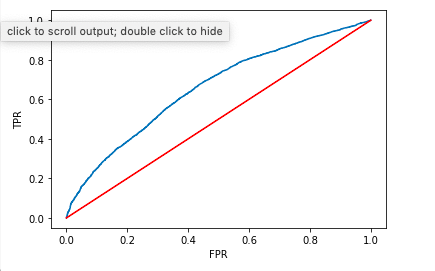




# 



# 



# 

# 

# 

# Summary and Conclusion

In this project we analyzed the given online news data set and were able to clearly observe some interesting patterns that good articles do have in common. We initially carried out a subjective analysis which was based on our own intuition and because we understand it is easily possible for human intuition to be biased or crowded from past experience, and use a quantitative analysis to confirm our initial hypothesis by doing univariate analysis using scatter plot, boxplot, and barplot of each feature with the shares feature.

Although our main gain is to build important insight about how popularity of articles are defined, we also went ahead in seeing how to predict the popularity of an article. Seven popularity classes were derived from the shared class and three machine learning models were built to be able to predict the popularity of the model. In order to be able to tune the models for better performances we consider different feature selection techniques, but these feature selection techniques didn’t really have much influence on the performance of the machine learning models.

The best machine learning model was the **Random Forest** which was able to attain an accuracy of 51.4% on the testing data-set. Some of the reasons for this low accuracy score is as a result of the large variance in the data set and also the imbalance in the class distribution which drives the prediction models to be bias towards popularity classes with more articles.

From the insight analysis carried out on the data-set the following are some of the things we recommend to improve the popularity of an article:

* The number of words in the article should be less than 1500 words. The lesser the better.
* Article title shouldn’t be too long or too short. 6 – 17 words is the ideal number of words to have for titles.
* Articles should have a good amount of images. Between 1 – 40 images is great.
* Also having a couple of videos is also nice for article popularity, but not too much. The higher the lower the odds.
* Easy to read words helps to improve article popularity.
* The number of keywords in the metadata really influences the shares to a margin. The higher the value the better the shares chances. A value upward of 5 is recommended.
* Articles referencing popular articles have a higher chance of improving their own popularity.
* Increase the number of popular unique words in the article to increase the chances of having better popularity.
* Avoid the use of longer words in the articles.
* Best popular articles are usually posted on Mondays and Wednesday (and a bit of Tuesdays). Sundays and Saturdays (Weekends generally) are the worst days to publish an article.
* Articles that talk about current trends tend to have higher popularity.
* Increase the amount of subjectivity in the title and content.
* The "Business" and "Entertainment" channels are great for the best popularity. Coming in third position will be the "World" and/or "Tech" channels.